Putting Human Assessments of Machine Translation Systems in Order

Putting Human Assessments of Machine Translation Systems in Order

Adam Lopez
Human Language Technology Center of Excellence
Johns Hopkins University

Abstract

Human assessment is often considered the gold standard in evaluation of translation systems. But in order for the evaluation to be meaningful, the rankings obtained from human assessment must be consistent and repeatable. Recent analysis by Bojar et al. (2011) raised several concerns about the rankings derived from human assessments of English-Czech translation systems in the 2010 Workshop on Machine Translation. We extend their analysis to *all* of the ranking tasks from 2010 and 2011, and show through an extension of their reasoning that the ranking is naturally cast as an instance of finding the minimum feedback arc set in a tournament, a well-known NP-complete problem. All instances of this problem in the workshop data are efficiently solvable, but in some cases the rankings it produces are surprisingly different from the ones previously published. This leads to strong caveats and recommendations for both producers and consumers of these rankings.

1 Introduction

The value of machine translation depends on its utility to human users, either directly through their use of it, or indirectly through downstream tasks such as cross-lingual information extraction or retrieval. It is therefore essential to assess machine translation systems according to this utility, but there is a widespread perception that direct human assessment is costly, unreproducible, and difficult to interpret. Automatic metrics that predict human utility have therefore attracted substantial attention since they are at least cheap and reproducible given identical data conditions, though they are frequently and correctly criticized for low interpretability and correlation with true utility. Their use (and abuse) remains contentious.

The organizers of the annual Workshop on Machine Translation (WMT) have taken a strong stance in this debate, asserting the primacy of human evaluation. Every annual report of their findings since 2007 has included a variant of the following statement:

> *It is our contention that automatic measures are an imperfect substitute for human assessment of translation quality. Therefore, we define the manual evaluation to be primary, and use the human judgments to validate automatic metrics.* (Callison-Burch et al., 2011)

The workshop's human evaluation component has been gradually refined over several years, and as a consequence it has produced a fantastic collection of publicly available data consisting primarily of *pairwise judgements* of translation systems made by human assessors across a wide variety of languages and tasks. Despite superb effort in the collection of these assessments, less attention has been focused on the final product derived from them: a *totally-ordered ranking* of translation systems participating in each task. Many of the official workshop results depend crucially on this ranking, including the evaluation of both machine translation systems and automatic metrics. Considering the enormous costs and consequences of the ranking, it is important to ask: is the method of constructing it accurate? The number of possible rankings is combinatorially large—with at least ten systems (accounting for more than

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 1–9,
Montréal, Canada, June 7-8, 2012.

half the cases we analyzed) there are over three million possible rankings, and with at least twenty (occurring a few times), there are over 10^{18} possible rankings. Exceptional care is therefore required in producing the rankings.

Bojar et al. (2011) observed a number of discrepancies in the ranking of English-Czech systems from the 2010 workshop, making these questions ever more pressing. We extend their analysis in several ways.

1. We show, through a logical extension of their reasoning about flaws in the evaluation, that the final ranking can be naturally cast as an instance of the *minimal feedback arc set* problem, a well-known NP-Hard problem.

2. We analyze 25 tasks that were evaluated using pairwise assessments from human annotators in 2010 and 2011.

3. We produce new rankings for each of the tasks, which are in some cases surprisingly different from the published rankings.

4. We identify a new set of concerns about sources of error and uncertainty in the data.

2 Human Assessment as Pairwise Ranking

The workshop has conducted a variety of different manual evaluation tasks over the last several years, but its mainstay has been the *relative ranking* task. Assessors are presented with a source sentence followed by up to five translations, and are asked to rank the translations from best to worst, with ties allowed. Since it is usually infeasible to collect individual judgements for all sentences for all pairs of systems on each task, consecutive sequences of three sentences were randomly sampled from the test data, with each sentence in each sequence presented to the same annotator. Some samples were presented multiple times to the same assessor or to multiple assessors in order to measure intra- and inter-annotator agreement rates. Since there are often more than five systems participating in the campaign, the candidate translations are likewise sampled from a pool consisting of the machine translations *and a human reference translation*, which is included for quality control purposes. It is important to note that the algorithm used to compute the published final rankings included *all* of this data, including comparisons against the reference and the redundant assessments used to compute inter-annotator agreement.

JHU	1	JHU$\prec$BBN-COMBO
BBN-COMBO	2	JHU$\prec$RWTH
RWTH	3	JHU$\prec$RWTH-COMBO
RWTH-COMBO	3	JHU$\prec$CMU
CMU	4	BBN-COMBO$\prec$RWTH
		BBN-COMBO$\prec$RWTH-COMBO
		BBN-COMBO$\prec$CMU
		RWTH$\equiv$RWTH-COMBO
		RWTH$\prec$CMU
		RWTH-COMBO$\prec$CMU

Figure 1: Example human relative ranking of five systems (left) and the inferred pairwise rankings (right) on a single sentence from the WMT 2010 German-English campaign.

The raw data obtained from this process is a large set of assessments. Each assessment consists of a list of up to five systems (including the reference), and a partial or total ordering of the list. The relative ranking of each pair of systems contained in the list is then taken to be their pairwise ranking. Hence a single assessment of five systems yields ten implicit pairwise rankings, as illustrated in Figure 1.

3 From Pairwise to Total Ranking

Given these pairwise rankings, the question now becomes: how do we decide on a total ordering of the systems? In the WMT evaluation, this total ordering has two critical functions: it is published as the official ranking of the participating systems; and it is used as the ground truth against which automatic evaluation metrics are graded, using Spearman's rank correlation coefficient (without ties) as the measure of accuracy. Choosing a total order is non-trivial: there are $N!$ possible orderings of N systems. Even with relatively small N of the workshop, this number can grow extremely large (over 10^{25} in the worst case of 25 systems).

The method used to generate the published rankings is simple. For each system A among the set S of ranked systems (which includes the reference),

compute the number of times that A is ranked better than or equivalent to *any* system $B \in S$, and then divide by the total number of comparisons involving A, yielding the following statistic for system A, which we call WMT-OFFICAL.

$$score(A) = \frac{\sum_{B \in S} count(A \preceq B)}{\sum_{B \in S, \diamond \in \{\prec, \equiv, \succ\}}, count(A \diamond B)} \quad (1)$$

The systems are ranked according to this statistic, with higher scores resulting in a better rank.

Bojar et al. (2011) raise many concerns about this method for ranking the systems. While we refer the reader to their paper for a detailed analysis, we focus on two issues here:

- Since ties are rewarded, systems may be unduly rewarded for merely being similar to others, rather than clearly better. This is of particular concern since there is often a cohort of very similar systems in the pool, such as those based on very similar techniques.
- Since the reference is overwhelmingly favored by the assessors, those systems that are more frequently compared against the reference in the random sample will be unfairly penalized.

These observations suggest that the statistic should be changed to reward only outright wins in pairwise comparisons, and to lessen the number of comparisons to the reference. While they do not recommend a specific sampling rate for comparisons against the reference, the logical conclusion of their reasoning is that it should not be sampled at all. This yields the following statistic similar to one reported in the appendices of the WMT proceedings, which we call HEURISTIC 2.

$$score(A) = \frac{\sum_{B \in S-ref} count(A \prec B)}{\sum_{B \in S-ref, \diamond \in \{\prec, \equiv, \succ\}}, count(A \diamond B)} \quad (2)$$

However, the analysis by Bojar et al. (2011) goes further and suggests disregarding the effect of ties altogether by removing them from the denominator. This yields their final recommended statistic, which we call BOJAR.

$$score(A) = \frac{\sum_{B \in S-ref} count(A \prec B)}{\sum_{B \in S-ref, \diamond \in \{\prec, \succ\}}, count(A \diamond B)} \quad (3)$$

Superficially, this appears to be an improvement. However, we observe in the rankings that two anonymized commercial systems, denoted ONLINEA and ONLINEB, consistently appear at or near the top of the rankings in all tasks. It is natural to wonder: even if we leave out the reference from comparisons, couldn't a system still be penalized simply by being compared against ONLINEA and ONLINEB more frequently than its competitors? On the other hand, couldn't a system be rewarded simply by being compared against a bad system more frequently than its competitors?

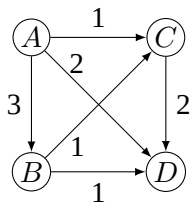
There are many possible decisions that we could make, each leading to a different ranking. However, there is a more fundamental problem: each of these heuristic scores is based on statistics aggregated over completely incomparable sets of data. Any total ordering of the systems must make a decision between every pair of systems. When that ranking is computed using scores computed with any of Equations 1 through 3, we aggregate over completely different sets of sentences, rates of comparison with other systems, and even annotators! Deriving statistical conclusions from such comparisons is at best suspect. If we want to rank A and B relative to each other, it would be more reliable to aggregate over the *same* set of sentences, *same* rates of comparison, and the *same* annotators. Fortunately, we have this data in abundance: it is the collection of pairwise judgements that we started with.

4 Pairwise Ranking as a Tournament

The human assessments are a classic example of a *tournament*. A tournament is a graph of N vertices with exactly $\binom{N}{2}$ directed edges—one between each pair of vertices. The edge connecting each pair of vertices A and B points to whichever vertex which is *worse* in an observed pairwise comparison between them. Tournaments are a natural representation of many ranking problems, including search results, transferable voting systems, and ranking of sports teams.[1]

Consider the simple weighted tournament depicted in Figure 2. This tournament is acyclic, which means that we can obtain a total ordering of the ver-

[1]The original motivating application was modeling the pecking order of chickens (Landau, 1951).



Consistent ranking: $A \prec B \prec C \prec D$

Ranking according to Eq. 1: $A \prec C \prec B \prec D$

Figure 2: A weighted tournament and two different rankings of its vertices.

tices that is consistent with all of the pairwise rankings simply by sorting the vertices topologically. We start by choosing the vertex with no incoming edges (i.e. the one that wins in all pairwise comparisons), place it at the top of the ranking, and remove it along with all of its outgoing edges from the graph. We then repeat the procedure with the remaining vertices in the graph, placing the next vertex behind the first one, and so on. The result is a ranking that preserves all of the pairwise rankings in the original graph.

This example also highlights a problem in Equation 1. Imagine an idealized case in which the consistent ranking of the vertices in Figure 2 is their true ranking, and furthermore that this ranking is unambiguous: that is, no matter how many times we sample the comparison A with B, the result is always that $A \prec B$, and likewise for all vertices. If the weights in this example represented the number of random samples for each system, then Equation 1 will give the inaccurate ranking shown, since it produces a score of $\frac{2}{5}$ for B and $\frac{2}{4}$ for C.

Tournaments can contain cycles, and as we will show this is often the case in the WMT data. When this happens, a reasonable solution is to minimize the discrepancy between the ranking and the observed data. We can do this by *reversing* a set of edges in the graph such that (1) the resulting graph is acyclic, and (2) the summed weights of the reversed edges is minimized. A set of edges satisfying these constraints is called the *minimum feedback arc set* (Figure 3).

The feedback arc set problem on general graphs

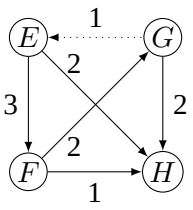


Figure 3: A tournament with a cycle on vertices E, F, and G. The dotted edge is the only element of a minimum feedback arc set: reversing it produces an acyclic graph.

Algorithm 1 Minimum feedback arc set solver

Input: Graph $\mathcal{G} = (V, E)$, weights $w : E \to \mathbb{R}^+$
Initialize all costs to ∞
Let $cost(\emptyset) \leftarrow 0$
Add $\emptyset$ to agenda $\mathcal{A}$
repeat
  Let $\hat{R} \leftarrow \text{argmin}_{R \in \mathcal{A}}\, cost(R)$
  Remove $\hat{R}$ from $\mathcal{A}$ ▷ $\hat{R}$ is a partial ranking
  Let $U \leftarrow V \backslash \hat{R}$ ▷ set of unranked vertices
  for each vertex $v \in U$ **do**
    Add $\hat{R} \cup v$ to agenda
    Let $c \leftarrow \sum_{v' \in U : \langle v', v \rangle \in E} w(\langle v', v \rangle)$
    Let $d \leftarrow cost(\hat{R}) + c$
    Let $cost(\hat{R} \cup \{v\}) \leftarrow \min(cost(\hat{R} \cup \{v\}), d)$
until $\text{argmin}_{R \in \mathcal{A}}\, cost(h) = V$

is one of the 21 classic problems shown to be NP-complete by Karp (1972).[2] Finding the minimum feedback arc set in a tournament was shown to be NP-hard by Alon (2006) and Charbit et al. (2007). However, the specific instances exhibited in the workshop data tend to have only a few cycles, so a relatively straightforward algorithm (formalized above for completeness) solves them exactly without much difficulty. The basic idea is to construct a dynamic program over the possible rankings. Each item in the dynamic program represents a ranking of some subset of the vertices. An item is extended by choosing one of the unranked vertices and appending it to the hypothesis, adding to its cost the weights of all edges from the other unranked vertices to the newly appended vertex (the

[2] Karp proved NP-completeness of the decision problem that asks whether there is a feedback arc set of size k; NP-hardness of the minimization problem follows.

Task name	#sys	#pairs	Task name	#sys	#pairs
2010 Czech-English	12	5375	2011 English-French individual	17	9086
2010 English-Czech	17	13538	2011 English-German syscomb	4	4374
2010 English-French	19	7962	2011 English-German individual	22	12996
2010 English-German	18	13694	2011 English-Spanish syscomb	4	5930
2010 English-Spanish	16	5174	2011 English-Spanish individual	15	11130
2010 French-English	24	8294	2011 French-English syscomb	6	3000
2010 German-English	25	10424	2011 French-English individual	18	6986
2010 Spanish-English	14	11307	2011 German-English syscomb	8	3844
2011 Czech-English syscomb	4	2602	2011 German-English individual	20	9079
2011 Czech-English individual	8	4922	2011 Spanish-English syscomb	6	4156
2011 English-Czech syscomb	2	2686	2011 Spanish-English individual	15	5652
2011 English-Czech individual	10	17875	2011 Urdu-English tunable metrics	8	6257
2011 English-French syscomb	2	880			

Table 1: The set of tasks we analyzed, including the number of participating systems (*excluding* the reference, #sys), and the number of implicit pairwise judgements collected (*including* the reference, #pairs).

edges to be reversed). This hypothesis space should be familiar to most machine translation researchers since it closely resembles the search space defined by a phrase-based translation model (Koehn, 2004). We use Dijkstra's algorithm (1959) to explore it efficiently; the complete algorithm is simply a generalization of the simple algorithm for acyclic tournaments described above.

5 Experiments and Analysis

We experimented with 25 relative ranking tasks produced by WMT 2010 (Callison-Burch et al., 2010) and WMT 2011 (Callison-Burch et al., 2011); the full set is shown in Table 1. For each task we considered four possible methods of ranking the data: sorting by any of Equation 1 through 3, and sorting consistent with reversal of a minimum feedback arc set (MFAS). To weight the edges for the latter approach, we simply used the difference in number of assessments preferring one system over the other; that is, an edge from A to B is weighted $count(A \prec B) - count(A \succ B)$. If this quantity is negative, there is instead an edge from B to A. The purpose of this simple weighting is to ensure a solution that minimizes the number of disagreements with all available evidence, counting each pairwise comparison as equal.[3]

[3]This is not necessarily the best choice of weighting. For instance, (Bojar et al., 2011) observe that human assessments of shorter sentences tend to be more consistent with each other, so perhaps they should be weighted more highly. Unfortunately, it is not clear how to evaluate alternative weighting schemes, since there is no ground truth for such meta-evaluations.

WMT-OFFICIAL (Eq 1)	MFAS	BOJAR (Eq 3)
ONLINE-B	CU-MARECEK	ONLINE-B
CU-BOJAR	ONLINE-B	CU-BOJAR
CU-MARECEK	CU-BOJAR	CU-MARECEK
CU-TAMCHYNA	CU-TAMCHYNA	CU-TAMCHYNA
UEDIN	CU-POPEL	CU-POPEL
CU-POPEL	UEDIN	UEDIN
COMMERCIAL2	COMMERCIAL1	COMMERCIAL2
COMMERCIAL1	COMMERCIAL2	COMMERCIAL1
JHU	JHU	JHU
CU-ZEMAN	CU-ZEMAN	CU-ZEMAN
38	0	69

Table 2: Different rankings of the 2011 Czech-English task. Only the MFAS ranking is acyclic with respect to pairwise judgements. The final row indicates the weight of the voilated edges.

An MFAS solution written in Python took only a few minutes to produce rankings for all 25 tasks on a 2.13 GHz Intel Core 2 Duo processor, demonstrating that it is completely feasible despite being theoretically intractible. One value of computing this solution is that it enables us to answer several questions,

ONLINEB	LIUM ≺ ONLINEB	1	RWTH-COMBO
RWTH-COMBO	UPV-COMBO ≺ CAMBRIDGE	6	CMU-HYPOSEL-COMBO
CMU-HYPOSEL-COMBO	JHU ≺ CAMBRIDGE	1	DCU-COMBO
CAMBRIDGE	LIMSI ≺ UEDIN	1	ONLINEB
LIUM	LIMSI ≺ CMU-HYPOSEL-COMBO	1	LIUM
DCU-COMBO	LIUM-COMBO ≺ CAMBRIDGE	1	CMU-HEAFIELD-COMBO
CMU-HEAFIELD-COMBO	LIUM-COMBO ≺ NRC	3	UPV-COMBO
UPV-COMBO	RALI ≺ UEDIN	1	NRC
NRC	RALI ≺ UPV-COMBO	4	CAMBRIDGE
UEDIN	RALI ≺ JHU	1	UEDIN
JHU	RALI ≺ LIUM	3	JHU-COMBO
LIMSI	LIG ≺ UEDIN	6	LIMSI
JHU-COMBO	BBN-COMBO ≺ NRC	3	RALI
LIUM-COMBO	BBN-COMBO ≺ UEDIN	5	LIUM-COMBO
RALI	BBN-COMBO ≺ UPV-COMBO	5	BBN-COMBO
LIG	BBN-COMBO ≺ JHU	4	JHU
BBN-COMBO	RWTH ≺ UPV-COMBO	3	RWTH
RWTH	CMU-STATXFER ≺ JHU	1	LIG
CMU-STATXFER	CMU-STATXFER ≺ LIG	1	ONLINEA
ONLINEA	ONLINEA ≺ RWTH	1	CMU-STATXFER
HUICONG	ONLINEA ≺ JHU	2	HUICONG
DFKI	HUICONG ≺ LIG	3	DFKI
CU-ZEMAN	DFKI ≺ RWTH	3	GENEVA
GENEVA	DFKI ≺ CMU-STATXFER	1	CU-ZEMAN

Table 3: 2010 French-English reranking with MFAS solver. The left column shows the optimal ranking, while the center shows the pairwise rankings that are violated by this ranking, along with their edge weights. The right column shows the ranking under WMT-OFFICIAL (Eq. 1), originally published as two separate tables.

both about the pairwise data itself, and the proposed heuristic ranking of Bojar et al. (2011).

5.1 Cycles in the Pairwise Rankings

Our first experiment checks for cycles in the tournaments. Only nine were acyclic, including all eight of the system combination tasks, each of which contained only a handful of systems. The most interesting, however, is the 2011 English-Czech individual task. This task is notable because the heuristic rankings *do not* produce a ranking that is consistent with all of the pairwise judgements, even though one exists. The three rankings are illustrated side-by-side in Table 2. One obvious problem is that neither heuristic score correctly identifies CU-MARECEK as the best system, even though it wins pairwise comparisons against all other systems (the WMT 2011 proceedings do identify it as a winner, despite not placing it in the highest rank).

On the other hand, the most difficult task to disentangle is the 2010 French-English task (Table 3), which included 25 systems (individual and system combinations were evaluated as a group for this task, despite being reported in separate tables in official results). Its optimal ranking with MFAS still violates 61 pairwise ranking samples — there is simply no sensible way to put these systems into a total order. On the other hand, the heuristic rankings based on Equations 1 through 3 violate even more comparisons: 107, 108, and 118, respectively. Once again we see a curious result in the top of the heuristic rankings, with system ONLINEB falling several spots below the top position in the heurstic ranking, despite losing out only to LIUM by one vote.

Our major concern, however, is that over half of the tasks included cycles of one form or another in the tournaments. This represents a strong inconsis-

tency in the data.

5.2 Evaluation of Heuristic Scores

Taking the analysis above further, we find that the total number of violations of pairwise preferences across all tasks stands at 396 for the MFAS solution, and at 1140, 1215, 979 for Equations 1 through 3. This empirically validates the suggestion by Bojar et al. (2011) to remove ties from both the numerator and denominator of the heuristic measure. On the other hand, despite the intuitive arguments in its favor, the empirical evidence does not strongly favor any of the heuristic measures, all of which are substantially worse than the MFAS solution.

In fact, HEURISTIC 2 (Eq. 2) fails quite spectacularly in one case: on the ranking of the systems produced by the tunable metrics task of WMT 2011 (Figure 4). Apart from producing a ranking very inconsistent with the pairwise judgements, it achieves a Spearman's rank correlation coefficent of 0.43 with the MFAS solution. By comparison, WMT-OFFICIAL (Eq. 1) produces the best ranking, with a correlation of 0.93 with the MFAS solution. The two heuristic measures obtain an even lower correlation of 0.19 with each other. This difference in the two rankings was noted in the WMT 2011 report; however comparison with the MFAS ranker suggests that the published rankings according to the official metric are about as accurate as those based on other heuristic metrics.

MFAS Ranking	HEURISTIC 2 Ranking
CMU-BLEU	CU-SEMPOS-BLEU
CMU-BLEU-SINGLE	NUS-TESLA-F
CU-SEMPOS-BLEU	CMU-BLEU
RWTH-CDER	CMU-BLEU-SINGLE
CMU-METEOR	STANFORD-DCP
STANFORD-DCP	CMU-METEOR
NUS-TESLA-F	RWTH-CDER
SHEFFIELD-ROSE	SHEFFIELD-ROSE

Table 4: Rankings of the WMT 2011 tunable metrics task. MFAS finds a near-optimal solution, violating only six judgements with reversals of CMU-METEOR $\prec$ CMU-BLEU and STANFORD-DCP $\prec$ CMU-BLEU-SINGLE. In contrast, the HEURISTIC2 (Eq. 2) solution violates 103 pairwise judgements.

6 Discussion

Unfortunately, reliably ranking translation systems based on human assessments appears to be a difficult task, and it is unclear that WMT has succeeded yet. Some results presented here, such as the complete inability to obtain a sensible ordering on the 2010 French-English task—or to produce an acyclic tournament on more than half the tasks—indicate that further work is needed, and we feel that the published results of the human assessment should be regarded with a healthy skepticism. There are many potential sources of uncertainty in the data:

- It is quite rare that one system is uniformly better than another. Rather, one system will tend to perform better in aggregate across many sentences. The number of sentences on which this improvement can be reliably observed will vary greatly. In many cases, it may be less than the number of samples.
- Individual assessors may be biased or malicious.
- The reliability of pairwise judgements varies with sentence length, as noted by Bojar et al. (2011).
- The pairwise judgements are not made directly, but inferred from a larger relative ranking.
- The pairwise judgements are not independent, since each sample consists of consecutive sentences from the same document. It is likely that some systems are systematically better or worse on particular documents.
- The pairwise judgements are not independent, since many of the assessments are intentionally repeated to assess intra- and inter-annotator agreement.
- Many of the systems will covary, since they are often based on the same underlying techniques and software.

How much does any one or all of these factors affect the final ranking? The technique described above does not even attempt to address this question. Indeed, modeling this kind of data still appears to be unsolved: a recent paper by Wauthier

and Jordan (2011) on modeling latent annotator bias presents one of the first attempts at solving just *one* of the above problems, let alone all of them.

Simple hypothesis testing of the type reported in the workshop results is simply inadequate to tease apart the many interacting effects in this type of data and may lead to many unjustified conclusions. The tables in the Appendix of Callison-Burch et al. (2011) report p-values of up to 1%, computed for every pairwise comparison in the dataset. However, there are over two thousand comparisons in this appendix, so even at an error rate of 1% we would expect more than twenty to be wrong. Making matters worse, many of the p-values are in fact much than higher than 1%. It is quite reasonable to assume that hundreds of the pairwise rankings inferred from these tables are incorrect, or at least meaningless. Methods for multiple hypothesis testing (Benjamini and Hochberg, 1995) should be explored.

In short, there is much work to be done. This paper has raised more questions than it answered, but we offer several recommendations.

- We recommend *against* using the metric proposed by Bojar et al. (2011). While their analysis is very insightful, their proposed heuristic metric is not substantially better than the metric used in the official rankings. If anything, an MFAS-based ranking should be preferred since it can minimize discrepancies with the pairwise rankings, but as we have discussed, we believe this is far from a complete solution.

- Reconsider the use of total ordering, especially for the evaluation of automatic metrics. As demonstrated in this paper, there are many possible ways to generate a total ordering, and the choice of one may be arbitrary. In some cases there may not be enough evidence to support a total ordering, or the evidence is contradictory, and committing to one may be a source of substantial noise in the gold standard for evaluating automatic metrics.

- Consider a pilot study to clearly identify which sources of uncertainty in the data affect the rankings and devise methods to account for it, which may involve redesigning the data collection protocol. The current approach is designed to collect data for a variety of different goals, including intra- and inter-annotator agreement, pairwise coverage, and maximum throughput. However, some of goals are at cross-purposes in that they make it more difficult to make reliable statistical inferences about any one aspect of the data. Additional care should be taken to minimize dependencies between the samples used to produce the final ranking.

- Encourage further detailed analysis of the existing datasets, perhaps through a shared task. The data that has been amassed so far through WMT is the best available resource for making progress on solving the difficult problem of producing reliable and *repeatable* human rankings of machine translation systems. However, this problem is not solved yet, and it will require sustained effort to make that progress.

Acknowledgements

Thanks to Ondřej Bojar, Philipp Koehn, and Martin Popel for very helpful discussion related to this work, the anonymous reviewers for detailed and helpful comments, and Chris Callison-Burch for encouraging this investigation and for many explanations and additional data from the workshop.

References

N. Alon. 2006. Ranking tournaments. *SIAM Journal on Discrete Mathematics*, 20(1):137–142.

Y. Benjamini and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300.

O. Bojar, M. Ercegovčević, M. Popel, and O. F. Zaidan. 2011. A grain of salt for the WMT manual evaluation. In *Proc. of WMT*.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proc. of WMT*.

C. Callison-Burch, P. Koehn, C. Monz, and O. F. Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proc. of WMT*.

P. Charbit, S. Thomass, and A. Yeo. 2007. The minimum feedback arc set problem is NP-hard for tournaments. *Combinatorics, Probability and Computing*, 16.

E. W. Dijkstra. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

R. M. Karp. 1972. Reducibility among combinatorial problems. In *Symposium on the Complexity of Computer Computations*.

P. Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proc. of AMTA*.

H. G. Landau. 1951. On dominance relations and the structure of animal societies: I effect of inherent characteristics. *Bulletin of Mathematical Biology*, 13(1):1–19.

F. L. Wauthier and M. I. Jordan. 2011. Bayesian bias mitigation for crowdsourcing. In *Proc. of NIPS*.